An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prevot, et al.

An empirical resource for discovering cognitive principles of discourse organization: the ANNODIS corpus

Stergos Afantenos[1], Nicholas Asher[1], Farah Benamara[1], Myriam Bras[2], Cécile Fabre[2], Mai Ho-dac[2], Anne Le Draoulec[2], Philippe Muller[1], Marie-Paule Péry-Woodley[2], Laurent Prévot[3], Josette Rebeyrolle[2], Ludovic Tanguy[2], Marianne Vergez-Couret[2], Laure Vieu[1]

[1]IRIT, Univ. Toulouse, France
[2]CLLE, Univ. Toulouse, France
[3]LPL, Univ. de Provence, France
(authors are in alphabetical order)

Abstract

here the abstract will go

1. Introduction

This paper describes the ANNODIS resource, a diversified corpus of written French texts enriched with several kinds of markup, including a manual annotation of discourse structures. The manual annotation is based on two approaches to discourse: a "bottom-up" approach whose aim is to construct the structure of a discourse from elementary units linked by coherence relations, and a "top-down" or "macro" approach which focuses on the selective annotation of multi-level discourse structures.

The ANNODIS corpus is the first such resource in French to our knowledge. But it also has distinct characteristics in comparison with English discourse annotated corpora like the Penn Discourse TreeBank or the RST tree bank. It is composed of texts that are diversified with respect to genre, length and type of discursive organization. It contains two distinct and complementary types of annotation. The bottom-up approach aims to provide a complete discourse structure for each text, starting from a segmentation of the text into elementary discourse units (EDUs), and then linking these by means of discourse relations, also known as coherence or rhetorical relations, to form complex discourse units or CDUs, which in turn may be linked via discourse relations to other discourse units. The top-down approach treats the document as a whole and seeks to find two types of high level structures that can also apply at more detailed levels—the so called "enumeration" structures and "topic chain" structures. The bottom-up approach exploits cues based on syntax, discourse markers and deep semantics, while the top-down approach exploits cues at the level of page layout as well as markers. The top-down approach provides a macro level organization that constrains the construction of CDUs in the bottom-up approach.

2. Choice of texts

The Annodis corpus is divided in two parts, corresponding to the two different approaches and annotation schemes. The bottom-up corpus consists of short texts (a few hundred words each) as the annotation process aims at a detailed analysis of every discourse unit. This annotation methods can also target excertps from longer documents. For the top-down approach, on the opposite, the annotation focuses on high-level discourse structures that appear at different levels of granularity and thus requires longer (several thousands words each), complete and more complex documents.

In order to provide a diversified corpus, we selected texts that show variations along three different characteristics: genre, type and document structure. Four different text genres are represented in the corpus, each issued from a different source: short news articles from the daily *Est Républicain*, encyclopedia articles (from the French Wikipedia), linguistics research papers (from *CMLF: Colloque Mondial de Linguistique Française*) and international relation reports (from *IFRI: Institut Français des Relations Internationales*). As for text types, we distinguish between narrative, expository and argumentative texts, each source providing a single text type. Finally, document structure is a rough measure of the amount of structuring features found in the documents (sections, headings, paragraphs, etc.) and is presented on a three level scale; here also this parameter is determined by the source.

Table 1, page 2, summarizes the content of the corpus, along with the number and total size of texts for each category. The first two rows describe the bottom-up part of the corpus, the last three the top-down part. However, there is some overlapping between these two subsets, as some of the top-down part have been annotated according to both methods, as presented in § 6..

Every text is protected by a Creative Commons license that allows us to make the Annodis corpus freely available for research purposes; this aspect played an important role in the selection of the sources.

3. Details on the Annotation Process

The ANNODIS resource provides two kinds of markups: rhetorical relations and multi-level discourse structures. Though the annotation of these markups is based on different approaches of discourse organisation (respectively bottom-up and top-down), different theoretical backgrounds and requires different types of text (see § 6.), the procedure is fairly similar: on the basis of an annotation

Id	Source	Genre	Type	Document structure	Texts	Tokens
NEWS	*Est Républicain*	news	narrative	low	39	10K
WIK1	*Wikipedia* excerpts	encyclopedia	expository	low	30	11K
WIK2	*Wikipedia*	encyclopedia	expository	high	30	231K
LING	*CMLF-08*	research	expository	medium	25	169K
GEOP	*IFRI (geo-political)*	reports	argumentative	medium	32	266K
Total					156	687K

Table 1: Breakup of the Annodis corpus

manual three naive coders annotated objects in texts by using a dedicated interface: Glozz (Mathet and Widlöcher, 2009). Glozz is an annotation tool, originally created for the annotation of the ANNODIS resource. This tool allows the annotation of units, relations and schemes plus a display of texts as real-life documents (with visual signalling such as paragraph breaks, headings, bullets/numbered lists, etc.) and a possibility for highlighting premarked features in order to assist annotation procedures. The next two subsections details the specific guidelines and give an overview of the data annotated for each kind of markups.

3.1. Bottom-Up Approach

The bottom-up approach used both naive and expert annotators. We have performed three phases of annotation. During the first preliminary phase two graduate-level students annotated 50 documents. We used their input in order to create an annotation manual which was used afterwards during the second, so called, "naive" phase. During this second phase 3 undergraduate students with no knowledge whatsoever of discourse theories doubly annotated 86 documents. The annotators were trained for a week, with the help of the aforementioned manual and the graphical annotation tool Glozz, designed to help them segment and annotate the documents as described in the previous section. During the last phase, expert annotators adjudicated the naive annotation on the 86 documents and corrected them.

The view of discourse structure underlying our approach is that common to RST (Mann and Thompson, 1987), LDM (Polanyi et al., 2004) the graphbank model (Wolf and Gibson, 2005), DLTAG (Forbes et al., 2003), PDTB (Prasad et al., 2008), and SDRT (Asher and Lascarides, 2003). SDRT served as the point of departure for the bottom-up annotation. Most of these theories define hierarchical structures by constructing CDUs from EDUs in recursive fashion. SDRT provides a graph-based view of discourse structure, which is more expressive than that of other theories (Danlos, 2007).

The relations linking DUs in this approach are a set of relations that are more or less common to all the theories of discourse mentioned above. We used earlier work on these relations and how they are linguistically marked to guide the annotation process. The linguistic marks include not only so-called discourse markers but also tense and aspectual shifts, as well as syntactic structure. The list of relations used is the following: EXPLANATION, GOAL, RESULT, PARALLEL, CONTRAST, CONTINUATION, ALTERNATION, ATTRIBUTION, BACKGROUND, FLASHBACK, FRAME, TEMPORAL-LOCATION, ELABORATION, ENTITY-ELABORATION, COMMENT.

Naive annotators were instructed to group EDU into complex units if these EDUs had a strong discursive unity and together play a discourse role.

	corpus total		**Est Rép**	**Wikip.**
Nb Textss	87		39	42
Nb words	28146		9768	17330
EDU	3188		1159	1949
CDU	1395		510	829
Discourse Relations				
	total (Nb)	**(%)**	**Est Républicain (%)**	**Wikipedia (%)**
alternation	18	0,5	0,3	0,6
attribution	75	2,2	3,0	1,7
background	155	4,6	5,2	4,8
comment	78	2,3	3,6	1,3
continuation	681	20,3	20,1	21,1
contrast	144	4,3	3,7	4,6
Eelab	527	15,7	14,1	16,4
elaboration	625	18,6	16,3	19,4
explanation	130	3,9	4,4	3,3
flashback	27	0,8	1,4	0,6
frame	211	6,3	6,2	5,7
goal	95	2,8	3,1	2,4
narration	349	10,4	11,1	10,4
parralel	59	1,8	2,2	1,8
result	163	4,9	4,7	5,4
temploc	18	0,5	0,5	0,5
totRel	3355	100	1203	2034

Table 2: Discourse relations of the expert annotations

3.2. Multi-level Structures annotation in a top-down approach

The top-down approach focuses on text organisation strategies and the detection of multi-level discourse structures (covering at least 2 sentences up to several headed-sections). The produced annotations concerned two multi-level discourse structures: topical chains and enumerative structures.

Topical chains (TCs) consist in a specific type of cohesive chain (Halliday and Hasan, 1976): topically homogeneous segments. These segments are mainly composed with connected units containing the same topical referent. The segments may contain sentences not topically connected to the others (e.g. comments, illustrations, etc.) if they occur between connected units as illustrated by the example given in Fig 1.

Enumerative structures (ESs) are segments of text resulting from the textual act of packaging and organizing independent elements according to an interpretative criterion following the definition given in (Luc et al., 2000):

> "The textual act [of Enumerating] consists in transposing textually the co-enumerability of the listed entities into the co-enumerability of the linguistic segments describing them, which thereby

Le LAF, rédigé en collaboration avec Igor Mel'cuk, est un travail qui a déjà mentionné à la section 4.1. En tant qu'ouvrage publié, **il** tire son originalité du fait qu'**il** est à la fois un manuel de lexicologie destiné, en tout premier lieu, aux enseignants de langue et un échantillon de dictionnaire du français, reposant sur une adaptation des descriptions formalisées de la LEC. **Il** s'accompagne d'un site web, où sont notamment rendus disponibles pour les enseignants de français des modèles d'exercices visant l'apprentissage du vocabulaire. Par sa finalité et par sa double nature (présentation de notions lexicologiques et de descriptions lexicographiques), **le LAF** peut être rapproché de Picoche (2007). Il est intéressant de constater que le travail d'interfaçage des principes et descriptions de la LEC opéré lors de la rédaction **du LAF** a permis, de façon rétroactive, de faire progresser l'approche théorique elle-même. On trouvera un bilan de l'expérience acquise au cours de la rédaction **du LAF** dans Polguère (2007). *Dans ce texte, on fait notamment état des innovations introduites pour ce qui est de la caractérisation sémantique des unités lexicales (au moyen d'étiquettes sémantiques) et de l'encodage des relations lexicales paradigmatiques et syntagmatiques (au moyen de formules dites « de vulgarisation »).* Une autre caractéristique originale **du LAF** est sa méthodologie d'élaboration (Polguère, 2000b). **Il** est en effet entièrement dérivé de la base lexicale DiCo des dérivations sémantiques et collocations du français, développée par Igor Mel'cuk et le présent auteur. Cette façon de procéder assure **au LAF** une rigueur formelle sous-jacente et, surtout, nous permet de dériver de la base source DiCo d'autres « produits », comme celui dont il va maintenant être question.	TC

Figure 1: TC – Topical Chain – example covering 2 paragraphs and mainly composed with connected units containing topical expressions referring to *Le LAF*. Sentence in italics is not about *Le LAF* but still included in the TC. Topical expressions are in bold.

II) Des orientations d'action Les orientations proposées peuvent être regroupées autour de quatre thèmes . - Mieux organiser notre politique étrangère dans la région ce qui passe, notamment, par la mise en place de structures permettant [...]. - Accentuer notre coopération avec des partenaires d'influence, notamment en établissant une coopération renforcée avec certains [...]. - Manifester notre souci de voir émerger des systèmes démocratiques dans la région en développant une politique d'influence auprès des [...]. - Contribuer plus efficacement à la solution des principales crises régionales, ce qui comporterait les actions suivantes : [...]. En conclusion, les turbulences qui affectent le moyen orient ont atteint un niveau de haute intensité qui représente, pour les pays occidentaux et, plus spécialement, pour l'Europe, de grands risques, notamment [...].	ES	TRIGGER ITEM 1 ITEM 2 ITEM 3 ITEM 4 CLOSURE

Figure 2: ES – Enumerative Structure – example covering a whole subsection and internally organised as follow: first, the heading together with the opening paragraph announce that the following text will list four themes of directions for action (re. the relationship between France and the Middle East); next, four bulleted items detail each of these theme, which are thereby presented as co-enumerable, i.e. identical in status with regard to the co-enumerability criterion; finally, the last paragraph of the subsection closes the enumeration with a conclusion.

> become the entities constituting the enumeration (the items). The identity of status of the items in the enumeration expresses the identity of status of the listed entities in the world". (Luc et al., 2000, p 25, our translation).

Around the enumeration as defined here, two optional segments may be found: a trigger and a closure. As a result, enumerative structures are characterised by an internal organisation involving 3 kinds of sub-segments: an optional **trigger** announcing the enumeration; several **items** composing the enumeration (at least two items must be identified for a structure to be present); an optional **closure** which summarises and/or closes the enumeration. Moreover, lexical expressions specifying the co-enumerability criteria may occur in the ES segment (more often in the trigger and/or the closure). In the ES' example given in Fig 2, "*thèmes*" is such an expression. We call such lexical expressions *enumeraTheme*.

The annotation of these two multi-level structures is detailed in an annotation manual produced to guide annotators. It distinguishes two stages: (1) identification of multi-level structures and delimiting segments (TCs and ESs) and sub-segments (triggers, items, closures) ; and (2) identification of the features signalling these structures (topical cues and trigger/item/closure cues).

Prior to annotation, a morphological and syntactic analysis was performed using TreeTagger and SYNTEX (Bourigault, 2007) which was used during the annotation procedure in order help annotators identify the desired structures and the features signalling them. The wide range of premarked features includes visual devices and document structure such as headings, bulleted/numbered items (Power et al., 2003; Péry-Woodley and Scott, 2006); punctuation (e.g. paragraphs ending with [:], punctuational motifs such as [: ...; ...; and/or ...]); and lexico-syntactic features based on studies about the signalling of discourse organisation. These lexico-syntactic features comprise coreferential and topical expressions (Cornish, 1999; Grosz et al., 1995; Gundel, 1998) e.g. pronouns and lexicla reiterations; item introducers (Turco and Coltier, 1988; Jackiewicz, 2005; Hempel and Degand, 2008) e.g. *firstly*, *finally*, *the first X*, *on the other hand,* ; predictive elements and anaphoric encapsulation (Francis, 1994; Bras et al., 2008; Legallois, 2006) ; sentence-initial circumstancial adverbials (as potential frame introducers (Charolles, 1997; Charolles M. et al., 2005)) ; other sentence-initial elements (e.g. connectives, appositions, etc.).

The annotation procedure processes as follow: once the text loaded into the interface, coders detect ESs and TCs by scanning the text with the help of visual layout and highlighting premarked features. Once a structure detected, they delimit the boundaries of each segments and sub-segments and, in the case of ES, the enumeraTheme i.e. the expression referring to the co-enumerability criterion. Hence, they identify features signalling these (sub-)segments by validating any pre-marked feature seen as relevant as well as identifying additional features that had not been pre-marked (such as syntactic parallelism, trigger reiteration).

The annotation was organized in three phases. During the first phase three texts were annotated by three annotators which could solicit expert annotators in order to resolve misunderstandings concerning the manual. After that, a second phase concerns the annotation of six texts by the 3 annotators. These first 27 annotated texts were used to measure the inter-annotator agreement in terms of F-measure which was 0.7 for ESs and 0.65 for TCs (calculated by comparing boundaries and cues indentification). These 27 texts

were also post-annotated in order to produce a gold version of them. Considering the F-measures as acceptable, the last phase of the annotation proceeded with the annotation of 73 texts by 1 annotator per text.

Combining these 3 phases, 1316 multi-level structures was annotated in 82 texts[1] (829 ESs and 487 TCs). Table 3 give a quantitative overview of the results of the annotation campaign, in terms of the different objects presented above and the different sub-corpus presented in § 2.:

corpus	ES	item	trigger	closure	enumeraTheme	TC
WIK2	332	1639	296	34	167	232
LING	263	838	224	46	151	68
GEOP	234	716	180	43	120	187
ANNODIS	829	3193	700	123	438	487

corpus	added features	validated premarked features
WIK2	1677	2428
LING	937	708
GEOP	1130	993
ANNODIS	3744	4129

Table 3: A quantitative overview of Multi-level Structures annotated

4. Some Experimentation and Future Work

4.1. EDU segmentation

We cast the task of EDU identification as a classification problem for each token, which can either start or end a DU, be a DU by itself, or be strictly contained within a DU.

For our classifier, we used a regularized maximum entropy model. The classification was followed by a post-processing enforcing well-balanced segments. After post-processing we had an F-measure of 0.733 for the EDUs as a whole. We present more details in (Afantenos et al., 2010).

4.2. Determining attachment points and the right frontier constraint

The right frontier constraint (RFC) in SDRT postulates that an incoming discourse unit should attach either to the last discourse unit or to one that is super-ordinate to it via a series of subordinate relations and complex segments (Asher and Lascarides, 2003). This postulate was never validated empirically at a corpus level. We used the Annodis data from the "naive" phase in order to check its validity. We found that the naive annotators, which had not been given any information on the structural postulates of SDRT, have respected the RFC in 95% of the cases. The 5% remaining was mostly annotation errors due to the fact that the graphical tool used was not well adapted for this task. More details are in (Afantenos and Asher, 2010). One practical implication is that the RFC can drastically reduce the search space for a discourse attachment, since we can consider as open to attachment only the nodes that are found on the RF.

[1]By taking into account the gold annotations rather than the annotations produced during the two first phases.

5. Results on multi-level structures annotations

5.1. Two frequent and well-identified textual strategies

Results of the annotation of high-level structures clearly establish that we are dealing with patterns of discourse organization that are intuitive and quite easy to annotate, as indicated by the good F-measures (3.2.). They are also very frequent, and they occur at different levels of the text structure, indicating that they are relevant patterns for studying the complexity of discursive organization. All three subcorpora in the ANNODIS corpus comprise a large number of these structures: from 5 to 12 topical chains per 10000 words, and from 11 to 18 enumerative structures. Topical chains occupy 15% of the text surface, enumerative structures 43%. Enumerative structures appear at different levels of granularity: each level of the text structure is concerned. They can stretch over several sections, several paragraphs, or they can occur within the limits of the paragraph. As for topical chains, the annotation programme limits the annotation to segments covering no more than one section (Fig 1 shows a one section TC). As a consequence, very high-level topcial chains was not annotated. These results show that both structures are a basic strategy to which writers resort frequently in different genres of expository texts. The following subsections focus on further results concerning enumerative structures (ESs).

5.2. A formal Typology of enumerative structures

A visual typology of enumerative structures has been proposed on account of their interaction with document structure at the different granularity levels that we have just mentioned. Type 1 are multisections ESs, where each item corresponds to a section (or subsection). Type 2 ESs are formatted lists. They are defined solely in terms of specific typographical and layout features (bullet points or numbers). They can be very local formatted lists composed of only two items or large-scale lists of up to 48 items covering an entire section. Type 3 ESs are multiparagraphic structures. On the most local level, type 4 depicts ESs that are inserted inside a paragraph or corresponding exactly to a paragraph. Concerning the main characteristics of these four visual types of ESs, some simple statistical measures provide the following interesting significant correlations: Types 1 and 2 are characterised by a higher cardinality (3.8 items on average against 3) and a higher presence of triggers; enumerathemes are more often present in Type 2 ESs and less often in Type 1 ESs; closures are significantly less frequent in Type 1 ESs. Cross-corpus comparisons are shown on table 4. These figures show that significant differences appear between corpora. Wikipedia articles are characterized by a larger amount of type 1 and particularly type 2 ESs, whereas local ESs are particularly present in the other two corpora, which resort less to multisection ESs.

5.3. Towards a functional typology of ES

As stated in 3.2., each ES may be associated by coders to lexical expressions referring to its co-enumerability criterion, what we have called 'enumeraTheme'. A first typology of annotations distinguishes three types: a concept (as

Corpus	Types of ESs			
	Headed sections	Formatted lists	Multi-paragraphic ESs	intra-paragraphic ESs
WIK1	19,3%	39,1%	20,8%	20,8%
LING	9,1%	23,2%	26,6%	41,1%
GEOP	6,8%	10,3%	20,9%	62%

Table 4: **Distribution of ESs types**

in 'the theory is based on three principles'), an entity (as in 'individuals are split up into 3 groups') or a textual object (as in 'this paper consists of four sections'). The vast majority (80%) of ESs concern concepts, against 9% of entities and 7.5% of textual objects. The 'concept' class must be refined, but this preliminary result suggests that ESs are predominantly creating new categories in discourse rather than making use of pre-existing categories.

6. Intersecting the bottom-up and top-down approaches and futur works

Given the top-down approach's hypothesis that high level structures affect the interpretation of other structures within their scope, we expect that top-down annotated structures will place constraints on the graph constructed via the bottom up method. Extracts of a subset of the texts in the WIK2, LIN AND GEO parts of the corpus were subject to both top-down and bottom up annotation methods, see table 5.

sub-corpus	Nb texts	Nb excerpts	N words
WIK2	9	12	4908
LING	3	3	1116
GEOP	3	3	1340

Table 5: Part of the ANNODIS corpus at the intersection of the two approaches

While a full understanding of the constraints induced by high level structures remains something for future study, several hypotheses already seem promising. 1) the macro-level structures can serve to guide CDU construction. As CDUs do not overlap, we predict that there should be no CDU that does not properly cover CDUs isolated by macro-methods. 2) macro-level structures such as enumerations can determine the semantic value of certain discourse markers like *puis*. If the overall structure, for instance, enumerates arguments in support of some hypothesis, a use of *puis* in the enumeration of those arguments should only be taken as indicating an instance of one of the arguments in the list, not a temporal sequence (which is what *puis* is typically used to do in the bottom-up approach). We hope to study constraints like these and enlarge the coverage of the doubly annotated corpus in future work.

7. Evaluating agreement

Evaluating agreement on complex relational data such as discourse annotations is far from obvious, and collecting this corpus has raised a number of interesting issues from this perspective. We focus here on the bottom-up case, which can be generalized to some of the top-down structures. Two kinds of information are annotated with a discourse graph: the attachment of discourse units to each other, and the labelling of the attachment arcs via discourse relations. We thus have two types of agreement to define, and the second one (relations labels) depend on the agreement for the first one (discourse unit pairs). One of our three annotators is much less in agreement with the other two than these between themselves, so we present the best correlated pair of annotators. We estimated the common proportion of attachments of one wrt the other as if the second one was the reference, which yields a F-score of 66%, for 279 common attachments. This is assuming attaching is a yes/no decision on every DU pair. But it should be noted this is not the way annotation works, as annotators try to cover minimally the text structure, and that some of these could be described in different syntactic ways, essentially with the use of complex units. The brutal estimation we give is thus likely to be an underestimation, and this raises the important issue of matching/comparing rhetorical structures. Refining this comparison is a work in progress. The agreement on labels was then computed on these commonly attached pairs, and yield a kappa of 0.4 for the full set of 17 relations. There is an important dispersion of annotations, and the majority class (entity elaboration) represents about 30% of the whole. We also evaluated agreement on groups of relations, for instance the groups of coordinating versus subordinating relations, similar to the distinction between satisfaction-precedence and hierarchical relations in (Grosz and Sidner, 1986), for which we got a kappa of .57. Again, this raises the issue of equivalent rhetorical structures which could be ascribed to the same portions of text, and we are working on defining a satisfactory discourse graph matching.

8. Conclusion

The ANNODIS corpus incorporates two levels of discourse annotation: a bottom-up type annotation of elementary and complex discourse units along with the coherence relations that connect those structures, and a top-down annotation of high level discourse structures such as enumerative structures. To our knowledge, this is the first such corpus for French but it also has several distinct characteristics that differentiate it from other more well-known resources in the English language. The bottom-up annotations of the ANNODIS corpus differ from those in the RST corpus, in that a wider array of structures are possible, and in that it distinguishes between complex discourse units and EDUs explicitly, which RST does not. Discourse pop-ups for non-contiguous spans of text are also explicitly marked. In relation to PDTB, the ANNODIS corpus creates full discourse structures instead of providing simply coherence relations between contiguous phrases. Finally, this corpus has led to the creation of various discourse-oriented tools (e.g., a segmenter) and has served to validate SDRT's right frontier constraint. The creation of a discourse parser is among our immediate goals as well.

9. References

Stergos D. Afantenos and Nicholas Asher. 2010. Testing SDRT's Right Frontier. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1–9.

Stergos D. Afantenos, Pascal Denis, Philippe Muller, and Laurence Danlos. 2010. Learning Recursive Segments for Discourse Parsing. In *Proceedings of LREC 2010*.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

D. Bourigault. 2007. Un analyseur syntaxique opérationnel : Syntex. Mémoire d'HDR, Université de Toulouse.

M. Bras, L. Prévot, and M. Vergez-Couret. 2008. Quelle(s) relation(s) de discours pour les structures énumératives ? In *Actes du Congrès Mondial de Linguistique Française*, pages 1945–1964, Paris.

Le Draoulec A. Charolles M., M.-P. Péry-Woodley, and L. Sarda. 2005. Temporal and spatial dimensions of discourse organisation. *Journal of French Language Studies*, 15(2):203–218.

Michel Charolles. 1997. L'encadrement du dicours : Univers, champs, domaines et espaces. *Cahier de Recherche Linguistique*, 6:1–73.

F. Cornish. 1999. *Anaphora, Discourse and Understanding. Evidence from English and French.* Calendron Press: Oxford.

Laurence Danlos. 2007. Strong generative capacity of RST, SDRT and discourse dependency DAGs. In A. Benz and P. Kühnlein, editors, *Constraints in Discourse*. Benjamins.

Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind K. Joshi, and Bonnie L. Webber. 2003. D-ltag system: Discourse parsing with a lexicalized tree-adjoining grammar. *Journal of Logic, Language and Information*, 12(3):261–279.

G. Francis. 1994. Labelling discourse: an aspect of nominal-group lexical cohesion. In M. Coulthard, editor, *Advances in Written Text Analysis*, pages 83–101. Routledge: London.

Barbara Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, July-September.

B.J. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

J. Gundel. 1998. Centering theory and the givenness hierarchy: Towards a synthesis. In A. Joshi M. Walker and E. Prince, editors, *Centering Theory in Discourse*, pages 183–198. Calendron Press: Oxford.

M.A.K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman: London.

Susanne Hempel and Liesbeth Degand. 2008. Sequencers in different text genres: Academic writing, journalese and fiction. *Journal of Pragmatics*, 40:676–693.

A. Jackiewicz. 2005. Les séries linéaires dans le discours. *Langue française*, 148:95–110.

D. Legallois. 2006. Quand le texte signale sa structure: La fonction textuelle des noms sous-spécifiés. *Corela*.

C. Luc, M. Mojahid, M.-P. Péry-Woodley, and J. Virbel. 2000. Les énumérations : structures visuelles, syntaxiques et rhétoriques. In *Actes de CIDE 2000 (Colloque International sur le Document Électronique)*, pages 21–40.

W. Mann and S. Thompson. 1987. Rhetorical structure theory : a theory of text organization. Technical report, Information Science Institute.

Y. Mathet and A. Widlöcher. 2009. La plate–forme GLOZZ : environnement d'annotation et d'exploration de corpus. In *TALN 2009*, Senlis, June. ATALA, LIPN.

M.-P. Péry-Woodley and D. Scott. 2006. Computational approaches to discourse and document processing. *TAL*, 2(47):7–19.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. A rule based approach to discourse parsing. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 108–117, Cambridge, Massachusetts, USA, April 30 - May 1. Association for Computational Linguistics.

R. Power, D. Scott, and N. Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 2(29):211–260.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

G. Turco and D. Coltier. 1988. Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire. *Pratiques*, 57:57–79.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus based study. *Computational Linguistics*, 31(2):249–287.